On the adaptation of the noise level for stochastic optimization

Olivier Teytaud, Anne Auger

On the adaptation of noise level for stochastic optimization

O. Teytaud and A. Auger

***Abstract*— This paper deals with the optimization of noisy fitness functions, where the noise level can be reduced by increasing the computational effort. We theoretically investigate the question of the control of the noise level. We analyse two different schemes for an adaptive control and prove sufficient conditions ensuring the existence of an homogeneous Markov chain, which is the first step to prove linear convergence when dealing with non-noisy fitness functions. We experimentally validate the relevance of the homogeneity criterion. Large-scale experiments conclude to the efficiency in a difficult framework.**

I. Introduction

Noise is present in many real-world optimization problems and can have various origins as measurement limitations or limited accuracy in simulation procedures. In general the precision of a fitness function evaluation depends on the computational effort (CE) and noise can be reduced by increasing the CE. For instance, the fitness evaluation can result from an expensive Monte-Carlo (quasi Monte-Carlo) simulation where the number of samples used for the simulation directly controls the precision of the accuracy of the evaluation. The fitness function may involve the resolution of Partial-Differential-Equation PDE where the control of the precision is driven by the resolution of the integration scheme or by POD-surrogate models (Proper Orthogonal Decomposition, [14]).

Since the noise level depends on the CE for lots of real-world situations, it is of major importance to understand how to control it in a sound way so as to guarantee good performances of optimization algorithms [8].

The terminology stochastic optimization, originally introduced to design the problem of solving noisy problems [21], [12], [9], [15], nowadays also refers to optimizing deterministic problems by means of stochastic algorithms. Among stochastic optimization algorithms, Evolutionary Algorithms (EAs) are well know to be suitable for optimizing real-world problems and to be in particular quite robust with respect to noise.

State-of-the-art EAs for continuous optimization are adaptive Evolution Strategies (ES), where the internal parameters of the mutation operator (standard deviation and covariance matrix) are adapted [11], [20], [19]. In practice the adaptation mechanisms allow to obtain linear convergence[1] for a wide class of uni-modal fitness functions. The question of linear convergence of adaptive ES can be theoretically addressed for simple fitness models by means of the theory of φ-irreducible Markov Chains [5], [3].

For noisy fitness functions with a constant noise amplitude (for instance Gaussian noise with constant standard deviation), EAs will fail in converging linearly to the optimum.

In this paper, we address the question of how to control the noise level and thus the CE to guaranty linear convergence. We focus on a simple adaptive $(1+1)$-ES on the noisy fitness model defined for $x \in \mathbb{R}^d$ as:

$$f_k(x,\eta) = \|x\|^k + \eta\mathcal{B} \text{ with } \|x\| = \sqrt{\sum_{i=1}^{d} x_i^2} \ , \qquad (1)$$

where $k \in \mathbb{R}^+\backslash\{0\}$, $\mathcal{B}$ is an independent random variable and $\eta \in \mathbb{R}^+$ is the parameter allowing to control the noise level. We present the first steps for the rigorous analysis of the linear convergence.

Note that reevaluating f_k allows to decrease the noise level: if $\mathcal{B}$ is a Gaussian white noise, *i.e.* $\mathcal{B} = \mathcal{N}(0,1)$ reevaluating n times f_k (*i.e* increase the computational effort by n) and average allows to decrease the noise level by a factor of $\sqrt{n}$.

In Section II we recall how Markov chains theory allows to prove linear convergence on the non-noisy version of Eq. (1) and present the homogeneous Markov Chain associated to the $(1+1)$-ES. In Section III-A we present a straightforward rule to adapt η that preserve this homogeneity, but unfortunately depends on some a priori information on the fitness. We then propose in Section III-B an alternate solution without prior knowledge of the fitness. In Section IV-A we present experiments on the fitness function 1 that backups our theoretical results. In Section IV-C we present experiments on a difficult fitness function.

II. Noise-free analysis: Markov chain and linear convergence

We consider, in this Section, an adaptive $(1 + 1)$-ES (Algorithm 1) minimizing $f : \mathbb{R}^d \to \mathbb{R}$. At each generation n, an offspring is sampled adding to the current parent x_n a Gaussian vector $\mathcal{N}_n$ with an identity covariance matrix scaled by a step-size σ_n (Line 6). At each generation, the step-size σ_n is increased in case of success (Line 11) and decreased otherwise (Line 15):

$$\sigma_{n+1} = \alpha\,\sigma_n \text{ if } f(x_n + \sigma_n\mathcal{N}_n) \leq f(x_n) \qquad (2)$$

$$= \beta\,\sigma_n \text{ otherwise.} \qquad (3)$$

with $\alpha > 1$ and $\beta < 0$. For isotropic ES, *i.e.* where the covariance matrix of the Gaussian vector is the identity, on the sphere function

$$f_s(x) = \|x\|^2,$$

TAO Team - INRIA Futurs, LRI, Univ. Paris-Sud 91405 Orsay, FRANCE olivier.teytaud@lri.fr, anne.auger@inria.fr

[1]Linear convergence means that the log of the distance to the optimum decreases linearly, for a fixed problem dimension, see Eq. (5) for the formal definition.

the optimal adaptation scheme for the step-size σ_n is

$$\sigma_n = \sigma_c \|x_n\| \tag{4}$$

where σ_c is a constant maximizing the (log)-progress rate (see [4] for instance). The $(1+1)$-ES implementing this adaptation scheme will converge linearly, *i.e.*

$$\exists\, \mathbf{c}(d) < 0 \text{ such that } \frac{1}{n} \ln \|x_n\| \to \mathbf{c}(d) \tag{5}$$

and the convergence rate $\mathbf{c}(d)$ reaches the lower bound for isotropic ES on the sphere [4], [23]. Of course the adaptation scheme given in Eq. (4) is artificial since it requires to know in advance the location of the optimum. The algorithm implementing this optimal adaptation scheme is scale-invariant on the sphere function. In particular, at each generation, the probability of success defined as the probability that one offspring is better than its parent is constant and roughly equal to $1/5$. From this $1/5$ factor, Rechenberg proposed the so-called one-fifth success rule for the adaptation of the step-size, aiming at maintaining a success probability of $1/5$ [19]. A robust implementation of the one-fifth success rule is to take $\alpha = 2$ and $\beta = 2^{-1/4}$ in Algorithm 1 [13].

Proving the linear convergence of the scale-invariant algorithm (where $\sigma_n = \sigma_c \|x_n\|$) stated in Eq. (5) is relatively easy [5]. However for "real" adaptation scheme like the one-fifth success rule the task is more complicated and calls upon the theory of φ-irreducible Markov Chains [17]. The basic steps for the analysis were pointed out in [5] for the analysis of self-adaptive ES and exploited in [3]. We recall here the main lines in the context of the adaptation scheme defined in Eqs. (2) and (3).

One iteration of Algorithm 1 can be summarized in the two equations:

$$x_{n+1} = x_n + \delta_n \sigma_n \mathcal{N}_n \tag{6}$$

$$\sigma_{n+1} = \sigma_n \mathcal{L}_n \tag{7}$$

where $\mathcal{N}_n$ is an independent Gaussian vector with covariance matrix identity, δ_n is a random variable equal to 1 whenever the offspring $x_n + \sigma_n \mathcal{N}_n$ is better than his parent and zero otherwise:

$$\begin{aligned} \delta_n = 1 \text{ if } & f_s(x_n + \sigma_n \mathcal{N}_n) \leq f_s(x_n) \\ = 0 \text{ otherwise.} & \end{aligned} \tag{8}$$

The random variable $\mathcal{L}_n$ is defined as

$$\begin{aligned} \mathcal{L}_n = \alpha \text{ if } & f_s(x_n + \sigma_n \mathcal{N}_n) \leq f_s(x_n) \\ = \beta \text{ otherwise.} & \end{aligned} \tag{9}$$

The proof of linear convergence of $(x_n)_{n\in\mathbb{N}}$ [3] relies on the fact that $(x_n/\sigma_n)_{n\in\mathbb{N}}$ is an homogeneous Markov chain:

Proposition 1 *Let x_n and σ_n be the random variables defined in Eq. (6) and Eq. (7). Then the sequence $z_n = x_n/\sigma_n$ is an homogeneous Markov chain. If z_n admits an invariant probability measure and is Harris recurrent, x_n converges linearly to zero, i.e. there exists a constant $\mathbf{c}_1$ such that:*

$$\frac{1}{n} \ln \|x_n\| \xrightarrow[n\to\infty]{} \mathbf{c}_1 \tag{10}$$

where $\mathbf{c}_1$ can be expressed in terms of the invariant measure of z_n.

Proof: First, we rewrite $\frac{x_{n+1}}{\sigma_{n+1}}$ as

$$\frac{x_{n+1}}{\sigma_{n+1}} = \frac{x_n + \delta_n \sigma_n \mathcal{N}_n}{\sigma_n \mathcal{L}_n} \tag{11}$$

$$= \frac{1}{\mathcal{L}_n} \frac{x_n}{\sigma_n} + \delta_n \frac{\mathcal{N}_n}{\mathcal{L}_n} \tag{12}$$

Besides, the offspring $x_n + \sigma_n \mathcal{N}_n$ is selected if:

$$\|x_n + \sigma_n \mathcal{N}_n\|^2 < \|x_n\|^2$$

But since σ_n is positive the sign of the previous equation is not changed if we divide it by σ_n, therefore the selection of the offspring will take place if:

$$\|x_n/\sigma_n + \mathcal{N}_n\|^2 < \|x_n/\sigma_n\|^2$$

Therefore the distribution of δ_n and $\mathcal{L}_n$ defined in Eqs. (8) and (9) does only depend on z_n. Eq. (12) can be rewritten as

$$z_{n+1} = \frac{1}{\mathcal{L}_n} z_n + \delta_n \frac{\mathcal{N}_n}{\mathcal{L}_n}.$$

We see that z_{n+1} does only depend on z_n (and not on previous iterate of the chain). Therefore z_n is a Markov chain. Besides there is no explicit dependence in n, implying that the Markov chain is homogeneous.

The second point of the proposition is to show that if the chain z_n is "stable", here positive and Harris recurrent, then linear convergence occurs, *i.e.* Eq. (10) is satisfied. First remark that

$$\frac{1}{n}(\ln \|x_n\| - \ln \|x_0\|) = \frac{1}{n} \sum_{k=0}^{n-1} \ln \|x_{k+1}\|/\|x_k\|$$

and that

$$\frac{\|x_{k+1}\|}{\|x_k\|} = \frac{\|z_{k+1}\|}{\|z_k\|} \mathcal{L}_k$$

Therefore

$$\frac{1}{n}(\ln \|x_n\| - \ln \|x_0\|) = \frac{1}{n} \sum_{k=1}^{n} \ln \|z_{k+1}\| \mathcal{L}_k/\|z_k\| \ .$$

If one were able to apply the Strong Law of Large Numbers (SLLN) to the right hand side of the previous equation, one would obtain Eq. (10). The Harris recurrence property and the existence of an invariant probability measure (positivity) are precisely the conditions required to be able to apply the SLLN. □

As we see in the previous Proposition, the first step to prove linear convergence is that x_n/σ_n is an homogeneous Markov chain. Proving the stability (existence of an invariant probability measure and Harris recurrence) is then a difficult second step, that can be addressed using drift conditions [3]. In this paper we will focus on the first step.

Note that since the selection in ES depends only on the ranking, Proposition 1 will hold for any $h \circ f_s$ where $h : \mathbb{R}^+ \to \mathbb{R}^+$ is any transformation preserving the rank (typically a monotonically strictly increasing function).

Algorithm 1 An adaptive $(1+1)$-ES.

Input: a fitness function f.
Initialize x_0, σ_0, $\alpha > 1$, $\beta < 1$.
Compute $fit_0 = f(x_0)$.
Initialize $n = 0$.
while true **do**
  $x'_n = x_n + \sigma_n \mathcal{N}_n$ [with $\mathcal{N}_n$ ind. isotropic Gaussian vector]
  Compute $fit = f(x'_n)$.
  if $fit < fit_n$ **then**
    $x_{n+1} = x'_n$
    $fit_{n+1} = fit$
    $\sigma_{n+1} = \alpha \sigma_n$ [Increase step-size]
  else
    $x_{n+1} = x_n$
    $fit_{n+1} = fit_n$
    $\sigma_{n+1} = \beta \sigma_n$ [Decrease step-size]
  end if
  $n \leftarrow n + 1$
end while

III. Noisy analysis

Our focus is on the extension of the baseline Algorithm 1 for the optimization of noisy functions, where the noise level is represented by a parameter η. We assume that for a prescribed noise level η we are able to adjust the computational effort associated to η: increase (resp. decrease) the CE to decrease (resp. increase) the noise level. The fitness model we consider for the theoretical analysis is defined in Eq. (1) and we investigate which adaptation rule for η does allow the $(1+1)$-ES to converge linearly. As a first step towards proving linear convergence we provide two schemes with underlying homogeneous Markov chains.

The noise parameter η being adapted, we denote η_n the noise level at iteration n. In [1], Arnold *et al.* use the progress rate approach to analyze the performance of the $(1+1)$-ES on the noisy sphere and they normalize the noise level by the distance to the optimum. In other words, η is adapted proportionately to the distance to the optimum, *i.e.* $\eta_n = \sigma_\epsilon \|x_n\|$ with $\sigma_\epsilon \in \mathbb{R}^+ \backslash \{0\}$. The first adaptation rule we study here is in the same vein but instead of considering the optimal adaptation rule for the step size, *i.e.* proportional to the norm, we consider a realistic update rule. In Section III-A η_n is proportional to the step-size σ_n, *i.e.* at each iteration n

$$\eta_n = (\sigma_n)^{k'} \tag{13}$$

where $k' \in \mathbb{R}^+$ is the tradeoff factor: k' larger leads to a larger computation time but a better precision. We show that $k' = k$ is a sufficient condition to have an homogeneous Markov chain on f_k.

The previous adaptation scheme depends on k and has no sense in discrete domains since it relies on σ_n. Therefore, we analyze in Section III-B the following rule

$$\eta_{n+1} = \mu \eta_n + \gamma(1-\mu)|fit_n - fit_{n-1}| \tag{14}$$

with $fit_{-1} = 0$, $\mu \in]0,1[$ and $\gamma > 0$.

A. Adaptation using step-size

In this section the noise level η is adapted at each generation using the step-size σ_n following Eq. (13). Let y_n denote the computed fitness at point x_n, *i.e.*

$$y_n = \|x_n\|^k + \sigma_{n'}^{k'} \mathcal{B}_{n'}$$

where $n' < n$ is the index of the last acceptance, $\mathcal{B}_{n'}$ the noise associated to this last acceptance. If $k' > 0$, this leads to more precise computations when the step-size is small.

Let b_n be the bias (or overvaluation [1]) of the evaluated fitness at iteration n, *i.e.*

$$b_n = y_n - \|x_n\|^k = \sigma_{n'}^{k'} \mathcal{B}_{n'} \tag{15}$$

As a first step towards proving linear convergence we try to identify an homogenous Markov chain associated to the algorithm and we show below that we can find an homogeneous Markov chain for $k' = k$. For the definitions below we consider that $k' = k$.

The $(1+1)$-ES can be summarized in the following equations:

$$\begin{aligned} x_{n+1} &= x_n + \delta_n \sigma_n \mathcal{N}_n \\ \sigma_{n+1} &= \sigma_n \mathcal{L}_n \\ b_{n+1} &= \delta_n \sigma_n^k \mathcal{B}_n + (1-\delta_n) b_n \end{aligned} \tag{16}$$

where

$$\begin{aligned} \mathcal{L}_n = \alpha \text{ if } \|x_n + \sigma_n \mathcal{N}_n\|^k + \sigma_n^k \mathcal{B}_n < \|x_n\|^k + b_n \\ = \beta \text{ otherwise} \end{aligned} \tag{17}$$

and

$$\begin{aligned} \delta_n = 1 \text{ if } \|x_n + \sigma_n \mathcal{N}_n\|^k + \sigma_n^k \mathcal{B}_n < \|x_n\|^k + b_n \\ = 0 \text{ otherwise.} \end{aligned} \tag{18}$$

Theorem 1 *Let x_n, σ_n and b_n be the sequences of random variables defined in Eq. (16), Eq. (17) and Eq. (18). Then*

$$\mathbf{Z_n} = \left(\frac{x_n}{\sigma_n}, \frac{b_n}{\sigma_n^k} \right).$$

is an homogeneous Markov chain. In other words for $k' = k$ the sequence of random variable $\mathbf{Z_n}$ induced by Algorithm 2 with $\eta_n = (\sigma_n)^{k'}$ is an homogeneous Markov chain.

Proof: Let denote $r_n = x_n/\sigma_n$ and $q_n = b_n/\sigma_n^k$. Then,

$$\begin{aligned} r_{n+1} = \frac{x_{n+1}}{\sigma_{n+1}} &= \frac{x_n + \delta_n \sigma_n \mathcal{N}_n}{\sigma_n \mathcal{L}_n} \\ &= \frac{1}{\mathcal{L}_n} \frac{x_n}{\sigma_n} + \delta_n \frac{\mathcal{N}_n}{\mathcal{L}_n} \\ &= \frac{1}{\mathcal{L}_n} r_n + \delta_n \frac{\mathcal{N}_n}{\mathcal{L}_n} \end{aligned}$$

$$\begin{aligned} q_{n+1} &= \frac{b_{n+1}}{\sigma_{n+1}^k} \\ &= \frac{\delta_n \mathcal{B}_n \sigma_n^k + (1-\delta_n) b_n}{(\mathcal{L}_n \sigma_n)^k} \\ &= \frac{1}{(\mathcal{L}_n)^k} \left(\delta_n \mathcal{B}_n + (1-\delta_n) q_n \right) \end{aligned}$$

Besides the selection step is determined by:

$$\|x_n + \sigma_n \mathcal{N}_n\|^k + \sigma_n^k \mathcal{B}_n < \|x_n\|^k + b_n$$

But since σ_n is positive the sign of the previous equation is not changed if we divide it by σ_n^k, therefore the selection step can be rewritten as:

$$\|r_n + \mathcal{N}_n\|^k + \mathcal{B}_n < \|r_n\|^k + q_n$$

Therefore $\mathbf{Z_{n+1}}$ only depends on $\mathbf{Z_n}$ and is defined as

$$r_{n+1} = \frac{1}{\mathcal{L}_n} r_n + \delta_n \frac{\mathcal{N}_n}{\mathcal{L}_n}$$
$$q_{n+1} = \frac{1}{(\mathcal{L}_n)^k} \left(\delta_n \mathcal{B}_n + (1-\delta_n) q_n\right)$$

with

$$\begin{aligned} \mathcal{L}_n = \alpha \text{ if } \|r_n + \mathcal{N}_n\|^k + \mathcal{B}_n < \|r_n\|^k + q_n \\ = \beta \text{ otherwise} \end{aligned} \tag{19}$$

and

$$\begin{aligned} \delta_n = 1 \text{ if } \|r_n + \mathcal{N}_n\|^k + \mathcal{B}_n < \|r_n\|^k + q_n \\ = 0 \text{ otherwise.} \end{aligned} \tag{20}$$

□

Algorithm 2 $(1+1)$-ES with adaptive rule for noise

Input: a noisy fitness function f.
Initialize x_0, σ_0, η, $\alpha > 1$, $\beta < 1$.
Compute $fit_0 = f(x_0, \eta)$.
Initialize $t = 0$.
while true **do**
  $\eta = \ldots$ **(eq. 13 or 14)** **[adaptation of η]**
  $x'_n = x_n + \sigma_n \mathcal{N}_n$ [with $\mathcal{N}_n$ ind. isotropic Gaussian vector]
  Compute $fit = f(x'_n)$.
  if $fit < fit_n$ **then**
    $x_{n+1} = x'_n$
    $fit_{n+1} = fit$
    $\sigma_{n+1} = \alpha \sigma_n$ [Increase step-size]
  else
    $x_{n+1} = x_n$
    $fit_{n+1} = fit_n$
    $\sigma_{n+1} = \beta \sigma_n$ [Decrease step-size]
  end if
  $n \leftarrow n + 1$
end while

B. An adaptive algorithm for noise

In this section the precision parameter η is chosen in a way that does not depend on the parameter k of the fitness. However we show that the result of Theorem 1 still holds.

As in the previous section y_n denotes the computed fitness at x_n. The $(1+1)$-ES can be summarized as follows:

$$\begin{aligned} x_{n+1} &= x_n + \delta_n \sigma_n \mathcal{N}_n \\ \sigma_{n+1} &= \sigma_n \mathcal{L}_n \\ y_{n+1} &= (1-\delta_n) y_n + \delta_n \|x_n\|^k + \delta_n \eta_n \mathcal{B}_n \end{aligned} \tag{21}$$

with

$$\begin{aligned} \delta_n = 1 \text{ if } \|x_n + \sigma_n \mathcal{N}_n\|^k + \eta_n \mathcal{B}_n < y_n \\ = 0 \text{ otherwise} \end{aligned} \tag{22}$$

and

$$\mathcal{L}_n = \delta_n \alpha + (1-\delta_n)\beta \tag{23}$$

The adaptation of the noise level η_n is done with the following rule

$$\eta_{n+1} = \mu \eta_n + \gamma(1-\mu)|y_{n+1} - y_n| \tag{24}$$

where $\mu \in]0,1[$ and $\gamma \in \mathbb{R}^+$.

Theorem 2 *Let $x_n, \sigma_n, y_n, \eta_n$ be defined by Eqs. (21), (22), (23), (24). Then*

$$\mathbf{Z_n} = \left(\frac{x_n}{\sigma_n}, \frac{y_n}{(\sigma_n)^k}, \frac{\eta_n}{(\sigma_n)^k}\right)$$

is an homogeneous Markov chain.

Proof: The evolution-equations can be rewritten as follows:

$$\begin{aligned} \frac{x_{n+1}}{\sigma_{n+1}} &= \frac{1}{\mathcal{L}_n}\left(\frac{x_n}{\sigma_n} + \delta_n \mathcal{N}_n\right) \\ \frac{y_{n+1}}{\sigma_{n+1}^k} &= (1-\delta_n)\frac{y_n}{\mathcal{L}_n^k \sigma_n^k} + \frac{\delta_n}{\mathcal{L}_n^k}\left(\frac{x_n^k}{\sigma_n^k} + \frac{\eta_n \mathcal{B}_n}{\sigma_n^k}\right) \\ \frac{\eta_{n+1}}{\sigma_{n+1}^k} &= \mu \frac{\eta_n}{\mathcal{L}_n^k \sigma_n^k} + \gamma(1-\mu)\left|\frac{y_{n+1}}{\sigma_{n+1}^k} - \frac{1}{\mathcal{L}_n^k}\frac{y_n}{\sigma_n^k}\right| \end{aligned}$$

and $\mathcal{L}_n$ only depends on δ_n, with δ_n only depending on $\|x_n\|^k + \eta_n \mathcal{B}_n < y_n$, i.e. only depending on $\|x_n\|^k/\sigma_n^k + \eta_n \mathcal{B}_n/\sigma_n^k$ and y_n/σ_n^k.

Therefore, $(x_{n+1}/\sigma_{n+1}, y_{n+1}/\sigma_{n+1}^k, \eta_{n+1}/\sigma_{n+1}^k)$ only depends on $(x_n/\sigma_n, y_n/\sigma_n^k, \eta_n/\sigma_n^k)$, $\mathcal{N}_n$ and $\mathcal{B}_n$ independently of n. This is a Markov chain and the transition does not depend on n: the Markov Chain is homogeneous. □

IV. Experiments

In this section we present experiments on the fitness function f_k (Section IV-A and IV-B) and on a fitness for the scrambling of quasi random sequences (Section IV-C).

In all the experiments, we took $\alpha = 2$ and $\beta = 2^{-1/4}$ implementing the one fifth-success rule [13]. The random variable $\mathcal{B}$ for the noise is uniform in $[0,1]$ and the initial point x_0 is drawn uniformly on the unit hypersphere.

Section IV-A shows experimentally that $\|x\|^2 + \sigma^{k'} B$ is solved linearly for $k' \geq 2$ (Theorem 1 states the homogeneity for $k' = 2$) and not solved at all for $k' < 2$. Section IV-B shows that the adaptive rule (14) works also whereas reducing the CE below the adaptive rule recommendation by an exponent < 1 does not work (this shows the optimality of our approach). Section IV-C shows an application to an important and difficult fitness function, namely the scrambling of quasi-random sequences.

A. Artificial experiments with fitness-specific precision Eq. (13)

Figure 1 shows that $(x_n)_{n\in\mathbb{N}}$ fails to converge lineraly on f_k as soon as $\eta_n = (\sigma_n)^{k'}$ with $k' < k$. For $k' = 2$ we observe a linear convergence. This suggests that the homogeneous Markov Chain of Theorem 1 is stable.

Figure 2 shows $\log(\|x_{11111}\|)$ on f_k for different values of k'. We see that $k' = k$ is optimal, suggesting that the homogeneity is a good criterion for choosing the noise level.

B. Artificial experiments with adaptive precision Eq. (14)

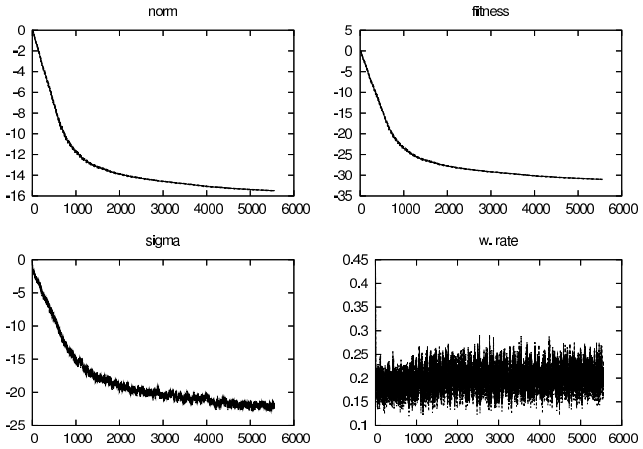
We consider now $\mathbf{f}_2 = \|x\|^2 + \eta^z\mathcal{B}$ in dimension 10 and Algorithm 2 with the adaptive rule given in (Eq. (14)). For $z = 1$, Theorem 2 states the homogeneity of the chain. In Fig 3, experiments for $z = 0.9$, $z = 1$ and $z = 1.1$ are shown for comparison. Once again, the homogeneity is emphasized as a good criterion for choosing the minimum CE for convergence to the optimum.

C. Minimization of L^2 discrepancy of a set of $10d^2$ points

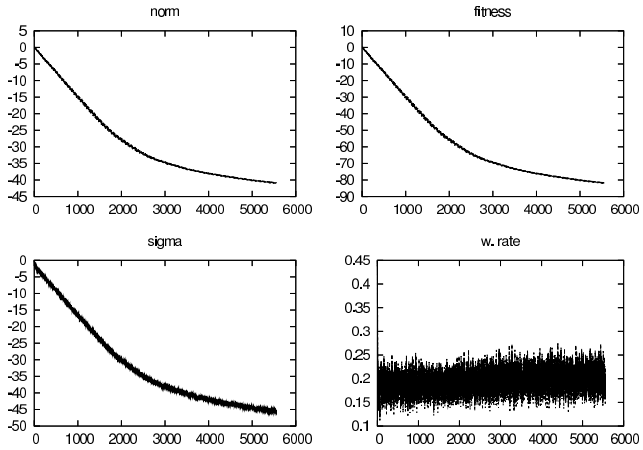
The theoretical analysis above is done in a continuous framework, with the noisy sphere-function, but Algorithm 2 with Eq. (14) used in Section IV-B can be used for discrete optimization also. Eq. (14) can indeed be used in any optimization algorithm (also non-evolutionary algorithms). The fitness investigated now is the L^2-discrepancy of a set of $10d^2$ points in dimension d generated thanks to a family of permutations; the domain is therefore a family of permutations (with some constraints that can be encoded in mutations). Several methods for choosing this family of permutations have been proposed in the literature: [6], [27], [7], [26], [18], [10], [29], [22], [28], [24], [2], [16]. All these solutions are analytically-designed thanks to mathematical analysis. In [25], a very efficient hand-designed solution, termed reverse-scrambling, is proposed. We here use a simple EA (Algorithm 3), with an approximated fitness by Monte-Carlo integration. Due to length constraints, we do not provide all details about the fitness; the interested reader is referred to [25] for all details.

We run this algorithm with $d = 3, 4, 5, 6, 7, 8, 9$ respectively, with $5d$ time-steps and $\mu = 0.9$. For each run, we compute the total computational cost and run the algorithm with η constant and the same overall computational cost. The results are presented in Table I. Note that the comparison with constant-CE is unfair in the sense that constant-CE has a prior information: the constant-CE benefits from the results performed by the CE-rule by using the average CE suggested by the CE-rule. Previous experiments for hand-tuning the noise level have been performed and it was a huge work; the success of the constant-CE-rule itself shows that the CE-level chosen by the CE-rule is a good one. On the other hand, the comparison with reverse scrambling is unfair; reverse scrambling is of course much faster (as it is analytically designed).

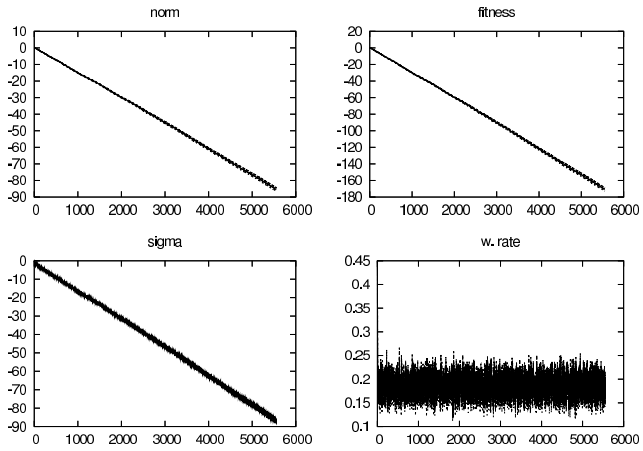
The success of our approach on this important problem shows the strong relevance of EA for such problems.



(a) k'=1.5



(b) k'=1.8



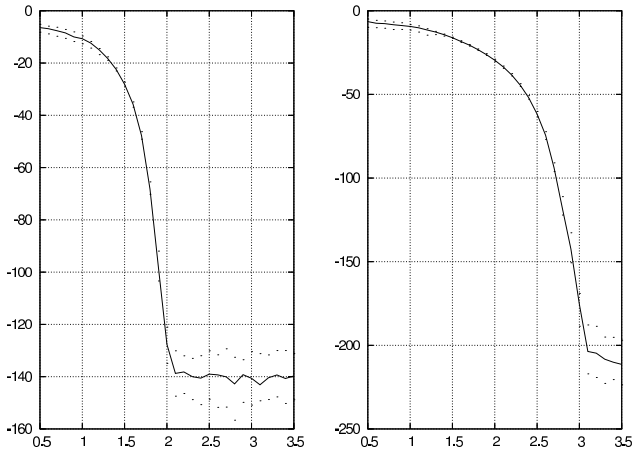
(c) k'=2

Fig. 1. Test on f_k using the adaptive rule (13): $\eta_n = (\sigma_n)^{k'}$ for $k' = 1.5, 1.8, 2$ in dimension 10. x-axis: number of iterations. Non-presented experiments show that the linear convergence is seemingly preserved for $k' \geq 2$. For each value of k, we present (clockwise) the log of the norm of x_n (i.e. a noise-free version of the fitness), the log of the fitness (including noise), the moving-average of the success rate, and $\log(\sigma)$. Dotted lines show the standard deviations.

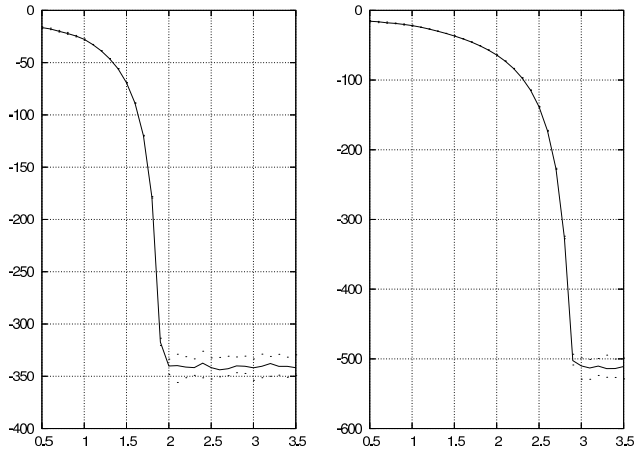
Method	L2-Discrepancy
Dimension 4	
Reverse Scr.	0.00714421 ± 1.01972e-05
RandomPoints	0.0155917 ± 0.000664881
Random Scr.	0.00778082 ± 0.000178898
CE-rule	0.00690391 ± 4.81381e-05
No scrambling	0.00833954 ± 1.2631e-05
constant-CE	0.00694316 ± 2.98309e-05
Dimension 5	
Reverse Scr.	0.00505218 ± 7.28368e-06
RandomPoints	0.00985676 ± 0.000452852
Random Scr.	0.00501275 ± 5.63656e-05
CE-rule	0.00468606 ± 3.34753e-05
No scrambling	0.00566526 ± 1.14062e-05
constant-CE	0.00468644 ± 3.20345e-05
Dimension 6	
Reverse Scr.	0.00325176 ± 4.94837e-06
RandomPoints	0.00638895 ± 0.000383821
Random Scr.	0.00355869 ± 4.18383e-05
CE-rule	0.00319954 ± 2.5507e-05
constant-CE	0.0032253 ± 1.7522e-05
No scrambling	0.00382715 ± 6.40585e-06
Dimension 7	
Reverse Scr.	0.00221943 ± 3.48892e-06
RandomPoints	0.00385059 ± 0.000164151
Random Scr.	0.00238406 ± 2.61057e-05
CE-rule	0.0022312 ± 7.87e-06
constant-CE	0.00225177 ± 1.20619e-05
No scrambling	0.00288602 ± 6.46388e-06
Dimension 8	
Reverse Scr.	0.00189975 ± 4.76932e-06
RandomPoints	0.00233921 ± 5.65993e-05
Random Scr.	0.00166693 ± 1.79506e-05
CE-rule	0.0015512 ± 1.28362e-05
No scrambling	0.00241595 ± 9.10999e-06
Dimension 9	
Reverse Scr.	0.00115382 ± 4.99428e-06
RandomPoints	0.00152759 ± 3.83705e-05
Random Scr.	0.00118502 ± 2.72275e-05
CE-rule	0.00106472 ± 4.15664e-06
No scrambling	0.00171113 ± 4.40706e-06

TABLE I

WE COMPARE HERE (I) REVERSE SCRAMBLING (II) RANDOM POINTS (III) RANDOM SCRAMBLING (IV) OUR CE-RULE (V) UNSCRAMBLED HALTON (VI) η CONSTANT AND OVERALL COST AS IN OUR CE-RULE. THE DIMENSIONALITY HERE REFERS TO THE DIMENSIONALITY OF THE UNDERLYING QUASI-MONTE-CARLO SEQUENCE AND NOT TO THE DIMENSIONALITY OF THE OPTIMIZATION PROBLEM. THE DOMAIN FOR DIMENSIONALITY d IS $E_1 \times E_2 \times \cdots \times E_d$, WHERE E_d IS THE SET OF PERMUTATIONS OF $[[1, p_d]]$ WHERE p_i IS THE i^{th} PRIME NUMBER. THE CONSTRAINTS ARE THAT 0 MUST BE FIXED POINT OF ALL PERMUTATIONS. IN ALL CASES, RANDOM SCRAMBLING WAS OUTPERFORMED BY THE CE-RULE, AND IN ALL BUT ONE CASE (DIM 7), REVERSE SCRAMBLING IS SIGNIFICANTLY OUTPERFORMED BY THE CE-RULE.



(a) dim=2



(b) dim=10

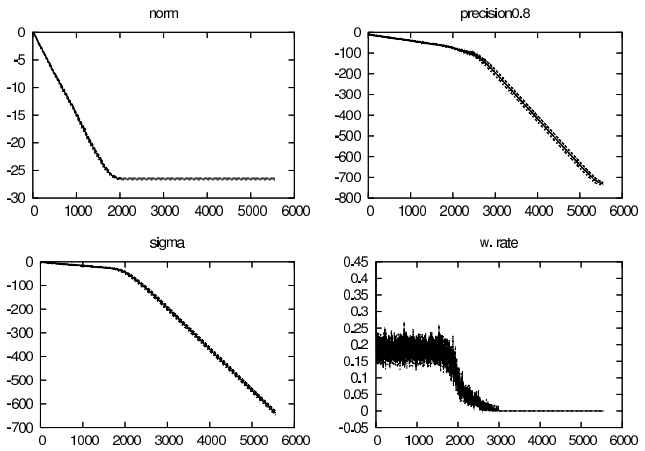
Fig. 2. Final log of fitness value at the 11111^{th} iterate in dimension 2 (top) and 10 (bottom) for various values of k' for the sphere function $\mathbf{f}_2 = \|x\|^2 + \eta\mathcal{B}$ (left) and $\mathbf{f}_3 = \|x\|^3 + \eta\mathcal{B}$ (right). x-axis: k'. y-values: final log-fitness. We see that $k' = 2 = k$ (left) and $k' = 3 = k$ (right) are the minimal possible choices ensuring convergence, as expected from theory. Dotted lines are the 10% and 90% percentiles.

V. CONCLUSION

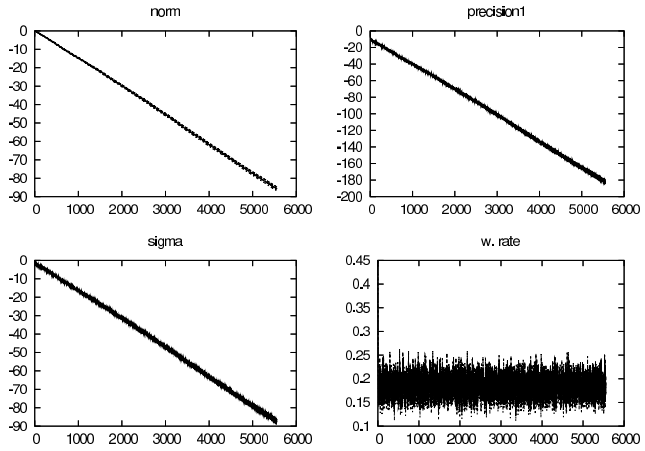
In this paper we have investigated the question of the adaptation of the noise level when optimizing noisy fitness function. This investigation is motivated by the fact that for many real-world problems the noise level can be reduced by increasing the computational effort. We have analyzed two different schemes. The first one is using the step-size σ_n of adaptive ES to control the noise level η, at iteration n:

$$\eta_n = (\sigma_n)^{k'}$$

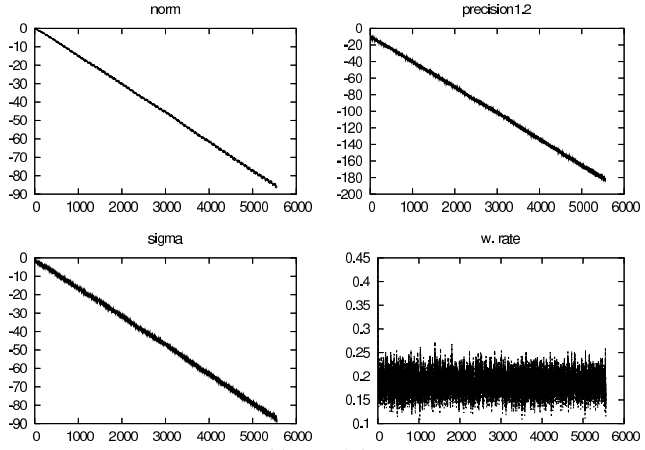
We have proved on the function $f_k(x) = \|x\|^k + \eta\mathcal{B}$, that $k' = k$ is a sufficient condition to have an homogeneous Markov Chains, first step to prove linear convergence when investigating non-noisy fitness functions. The experiments performed show that for $k' < k$, the algorithms fails to converge linearly and suggest that $k' = k$ is the optimal choice. Thus an adaptation scheme allowing linear convergence depends on the knowledge of the fitness function: the



(a) z=0.8



(b) $z = 1.0$



(c) $z = 1.2$

Fig. 3. Test with $z = 0.8$, $z = 1.0$, $z = 1.2$ in $f(x, \eta) = \|x\|^2 + \eta^z B$. Theory predicts a linear behavior for $z = 1$; we here verify on experiments that $z \geq 1$ is seemingly necessary for convergence. x-values: iterations. For each value of z, we present (clockwise) the log of the norm of x_n (i.e. a noise-free version of the fitness), the log of α (with title "precision$k$"), the moving-average of the success rate, and $\log(\sigma)$. Interestingly, for $z < 1$, we see that the algorithm is not only slow: it does not converge and $\sigma \to 0$ (until the machine precision) without further improvement of the fitness. On the other hand, increasing z, in spite of the larger CE, does not improve the result.

Algorithm 3 A $(1 + 1)$-EA with noise for scrambling-optimization.

Initialize x_0 (constant permutations, i.e. unscrambled Halton-sequence), η.
Compute $fit_0 = f(x_0, \eta)$.
Initialize $n = 0$.
while true **do**
  $\eta \leftarrow \mu\eta + \gamma(1 - \mu)(fit_n - fit_{n-1})$
  Set x'_n equal to x_n, plus some random transposition respecting the constraints.
  Compute $fit = f(x'_n, \eta)$.
  if $fit < fit_n$ **then**
    $x_{n+1} = x'_n$
    $fit_{n+1} = fit$
  else
    $x_{n+1} = x_n$
    $fit_{n+1} = fit_n$
  end if
  $n \leftarrow n + 1$
end while

factor k.

The second adaptation scheme investigated is independent of the knowledge of fitness function and is adaptive:

$$\eta_{n+1} = \mu\eta_n + \gamma(1 - \mu)|fit_n - fit_{n-1}|$$

We also prove the existence of an homogeneous Markov chain for this scheme. We apply this scheme to optimize an important and difficult fitness function, namely the scrambling of Quasi-Random sequences. The algorithm found the right level of CE in this very hard framework.

We presented the very first steps of the mathematical analysis of the linear convergence since we only exhibited homogeneous Markov chains. Our experiments confirmed the relevance of the homogeneity criterion for choosing the CE. Deriving the stability of the different Markov chains to prove the linear convergence is the object of further research.

The algorithms we propose are probably not the optimal possible ones. It is reasonnable to improve the precision of the current iterate, when too much time is spent on the same point, in particular for elitist strategies; this is not done in our work and could be done while preserving the homogeneity. This will be the object of a further analysis.

One final remark is that in the case of additive noise, one looses the invariance to order preserving transformations.

REFERENCES

[1] D. V. Arnold and H.-G. Beyer. Local performance of the (1+1)-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41, 2002.

[2] E. Atanassov. On the discrepancy of the halton sequences. *Math. Balkanica*, 18(12):1532, 2004.

[3] A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 334:35–69, 2005.

[4] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In A. Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[5] A. Bienvenüe and O. Fancois. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Theor. Comput. Sci.*, 306(1-3):269–289, 2003.

[6] E. Braaten and G. Weller. An improved low-discrepancy sequence for multidimensional quasi-monte carlo integration. *J. Comput. Phys.*, 33:249–258, 1979.

[7] R. Cranley and T. Patterson. Randomization of number theoretic methods for multiple integration. *SIAM J. Numer. Anal.*, 13(6):904914, 1976.

[8] J. Dennis and V. Torczon. Managing approximation models in optimization. In *In Alexandrov, N. and Hussaini, M. Y., editors, Multidisciplinary Design Optimization: State of the Art.*, 1996.

[9] B. Denton. Review of "stochastic optimization: Algorithms and applications" by Stanislav Uryasev and Panos M. Pardalos, Kluwer Academic Publishers 2001. *Interfaces*, 33(1):100–102, 2003.

[10] H. Faure. Good permutations for extreme discrepancy. *J. Number Theory*, 42:4756, 1992.

[11] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[12] P. Kall. *Stochastic Linear Programming*. Springer, Berlin, 1976.

[13] S. Kern, S. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos. Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review. *Natural Computing*, 3:77–112, 2004.

[14] F. Leibfritz and S. Volkwein. Reduced order output feedback control design for PDE systems using proper orthogonal decomposition and nonlinear semidefinite programming. *Linear Algebra and Its Applications*, 415:542–757, 2006.

[15] K. Marti. *Stochastic Optimization Methods*. Springer, 2005.

[16] M. Mascagni and H. Chi. On the scrambled halton sequence. *Monte Carlo Methods Appl.*, 10(3):435–442, 2004.

[17] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.

[18] W. Morokoff and R. Caflish. Quasi-random sequences and their discrepancies. *SIAM J. Sci. Comput.*, 15(6):12511279, 1994.

[19] I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.

[20] H.-P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1981. 1995 – 2^{nd} edition.

[21] J. K. Sengupta. *Stochastic Programming. Methods and Applications*. North-Holland, Amsterdam, 1972.

[22] A. Srinivasan. Parallel and distributed computing issues in pricing financial derivatives through quasi-monte carlo. In *Proceedings of the 16th International Parallel and Distributed Processing Symposium*, 2002.

[23] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, 2006.

[24] B. Tuffin. A new permutation choice in halton sequences. *Monte Carlo and Quasi-Monte Carlo*, 127:427435, 1997.

[25] B. Vandewoestyne and R. Cools. Good permutations for deterministic scrambled halton sequences in terms of l2-discrepancy. *Computational and Applied Mathematics*, 189(1,2):341:361, 2006.

[26] X. Wang and F. Hickernell. Randomized halton sequences. *Math. Comput. Modelling*, 32:887–899, 2000.

[27] T. Warnock. Computational investigations of low-discrepancy point sets. In *In: S.K. Zaremba, Editor, Applications of Number Theory to Numerical Analysis (Proceedings of the Symposium, University of Montreal*, page 319343, 1972.

[28] T. Warnock. Computational investigations of low-discrepancy point sets ii. In *In: H. Niederreiter and P.J.-S. Shiue, Editors, Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer, Berlin*, 1995.

[29] G. Okten and A. Srinivasan. Parallel quasi-monte carlo methods on a heterogeneous cluster. In *in: H. Niederreiter, K.-T. Fang, F.J. Hickernell (Eds.), Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer, Berlin, Heidelberg*, page 406421, 2002.